REPORT DOCUMENTATION PAGE

88

AFRL-SR-AR-TR-04-

0584

1. AGENCY USE ONLY *(Leave blank)*	2. REPORT DATE	01 May 2003 - 31 Jul 2004 FINAL

4. TITLE AND SUBTITLE
New Forcefields and Algorithms for Computational Protein Design and Docking

5. FUNDING NUMBERS
61101E
P956/00

6. AUTHOR(S)
Dr Sapiro

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)
UNIVERSITY OF MINNESOTA
200 OAK STREET SE
MINNEAPOLIS NM 55455-2070

8. PERFORMING ORGANIZATION REPORT NUMBER

9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)
AFOSR/NE
4015 WILSON BLVD
SUITE 713
ARLINGTON VA 22203

10. SPONSORING/MONITORING AGENCY REPORT NUMBER
F49620-03-1-0279

11. SUPPLEMENTARY NOTES

12a. DISTRIBUTION AVAILABILITY STATEMENT
DISTRIBUTION STATEMENT A: Unlimited

12b. DISTRIBUTION CODE

13. ABSTRACT *(Maximum 200 words)*

This seed project has achieved its short-term and long—term goals. In the short-term, we have developed novel techniques for protein-protein docking and for studying coiithrniation spaces. In the long-term, we have started a fruitful collaboration between a biochemist (Prof Baker) and a computational/theory researcher (Prof Sapiro). Prof Sapiro and two of his students (D. Rother and P. Lloyd) visited Prof Baker and his group a number of times (2-3 times each) during the year of this project.

20041129 019

14. SUBJECT TERMS			15. NUMBER OF PAGES
			16. PRICE CODE
17. SECURITY CLASSIFICATION OF REPORT	18. SECURITY CLASSIFICATION OF THIS PAGE	19. SECURITY CLASSIFICATION OF ABSTRACT	20. LIMITATION OF ABSTRACT
Unclassified	Unclassified	Unclassified	UL

Final Report - DARPA Grant F49620-03-1-0279
New Forcefields & Algorithms for Computational Protein Design

Guillermo Sapiro
University of Minnesota
guile@ece.umn.edu

This seed project has achieved its short-term and long–term goals. In the short-term, we have developed novel techniques for protein-protein docking and for studying conformation spaces. In the long-term, we have started a fruitful collaboration between a biochemist (Prof. Baker) and a computational/theory researcher (Prof. Sapiro). Prof. Sapiro and two of his students (D. Rother and P. Lloyd) visited Prof. Baker and his group a number of times (2-3 times each) during the year of this project.

We will now proceed to briefly describe the two main projects addressed during this one year grant.

The protein-protein docking problem, that is, the task of assembling two separate protein components into their biologically-relevant complex structure, is important for several reasons. First, it is of extreme relevance to cellular biology, where function is accomplished by proteins interacting with themselves and with other molecular components. Second, the protein docking problem presents one of the fundamental tests of the understanding of molecular biophysics, requiring a sophisticated knowledge of molecular motions and free energy calculations. Finally, an important post-genomic goal is the determination of the structural details of interactions between all pairs of proteins that bind, and computational tools offer an inexpensive means to prepare large-scale studies. Most currently available algorithms employ a docking procedure that is either based on the chemical and energetic properties of the protein molecules or on geometric and topological properties. These algorithms alone all have deficiencies in computing a successful dock, and for a docking algorithm to function correctly, a combination of these two types of techniques must be utilized. In this work we extend the contribution in Baker's group protein-protein docking procedure, which uses a very detailed representation of side-chain energetic and conformational freedom. Our contribution is in the form of a geometric and statistical based filter. This filter, which is imperative to improve the docking results and to reduce the computational cost demanded by such accurate representation, consists of a surface feature identification algorithm constructed on a geometric grid-based technique, along with a simple numeric comparison of proteins based on statistical measurements. The geometric component is based on a new algorithm for detecting and analyzing cavities, a technique that simulates water filling. This is combined with a novel technique for addressing the compactness of the dock. The combined filter has been able to eliminate over 70% of the decoys or false candidates produced by the accurate energetic computations, and shows great promise for the continued development of protein docking algorithms based on the combination of biolochemical with geometric criteria. An example of a dock filtered out by our technique, and that was not filtered out without the geometric and statistical additions, is

presented in Figure 1. A paper reporting these results is currently in preparation and will be submitted by the end of 2004.

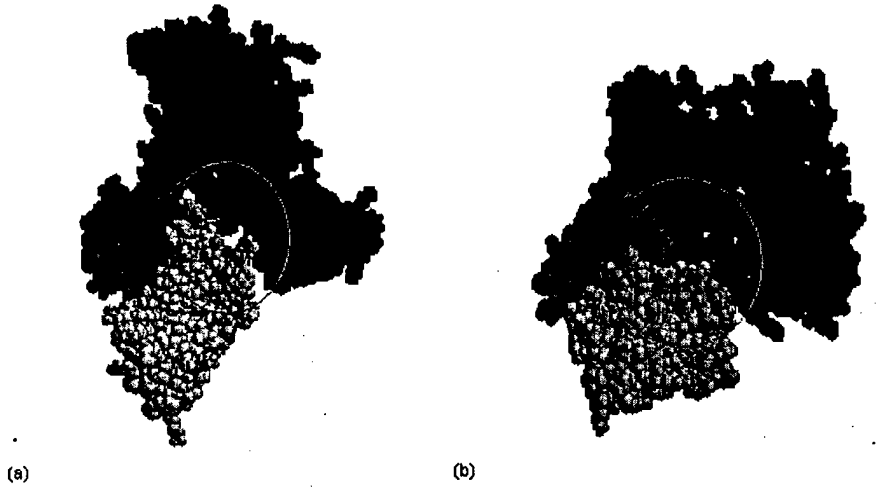

Figure 1. This figure illustrates the fact that while a docking algorithm based almost purely on protein energetics predicts a successful dock, the prediction may be inaccurate from a geometric standpoint, even if good complementarity scores are achieved. (a) A decoy/candidate from the target set 1EO8, which has a very high score based on Baker's scoring scheme, but does not dock very tightly due to inherent geometric flaws in the decoy's design. (b) Another decoy from target set 1EO8 representing a tighter fitting dock. Pure complementarity measures will also fail in detecting such small geometric problems, while the combination of geometric and statistical filters we have developed under this project do manage to distinguish between the two docks.

The second project we have addressed is the study of the space of conformations. From the physical measurements of protein conformations, e.g., via X-ray crystallography, as well as from molecular dynamic simulations as those produced by Prof. Pande's group at Stanford University, more than one conformation is obtained for a given protein. This multiple representations have not been sufficiently exploited in areas such as docking and design. We started in this direction by computing the space of conformations of a given protein. The challenge here is the estimation of a probability density function in high dimensional space and with very few samples. This is done with a combination of theories, including kernel smoothing and bootstrapping. Once these distributions are obtained, they can be used for a large number of applications, including high-resolution protein design, docking, and protein classification. A paper describing these results is currently in preparation and will be submitted by the end of 2004.